# COMPARATIVE STUDY OF THE GENERALIZED DFIIT STRUCTURE AND ITS EQUIVALENT STATE-SPACE REALIZATION

Z.X Zhao and G. Li

School of EEE
Nanyang Technological University
Singapore 639798

ABSTRACT

The generalized direct-form II transposed (DFIIt) structure in $\delta$-operator has been studied by several authors recently and it has been shown that this structure has some nice numerical properties over that in the conventional shift operator. Noting that the $\delta$-operator based DFIIt structure yields a better performance only for those digital IIR filters whose poles are clustered around $z = +1$, based on a polynomial operator approach a more generalized DFIIt structure is derived and its equivalent state-space realization is investigated in this paper. The corresponding expression for roundoff noise gain is derived. The optimal polynomial problems are then formulated and solved for any given filter. It is shown that this realization yields a smaller roundoff noise gain than that of the generalized DFIIt. A numerical example is given to illustrate the design procedure.

## 1. INTRODUCTION

The optimal structure design has been considered as one of the most effective methods (see, e.g., [1]-[4]) to minimize the effects of finite word length (FWL) errors on the performance of digital filters. It is well known that for a given digital filter, there exist a number of different realizations with which the filter can be implemented. The optimal FWL state-space design is to compute those realizations that minimize the degradation of the filter due to the FWL effects. It has been noted that the optimal realizations are usually fully parameterized. In practice, it is desired that the filter have a nice performance as well as a very simple structure that possesses many trivial parameters[1], which can be implemented exactly and produce no rounding errors. Noting this fact, a lot of effort has been made to achieve sparse optimal or quasi-optimal realizations (see, e.g., [5]-[6]).

It is well known that though having poor numerical properties the direct-forms in the conventional shift operator are the simplest structures. Recently, the direct-forms in delta operator have been studied by researchers (see, e.g., [7]-[11]). An extensive comparative study of different direct forms in delta operator was given in [11], where the transfer function is cascaded into second order sections and each section is implemented with a direct form in delta operator. It was shown there that among all the direct forms, the direct form II transposed (DFIIt) structure has the lowest quantization noise level at output. In [7], the DFIIt structure in $\delta$-operator was investigated for an arbitrary order IIR filter, where the concept of different coupling coefficients at different branch nodes is utilized for better roundoff noise gain suppression.

The use of delta operator, defined as $\delta = \frac{z-1}{T_s}$ with $T_s$ the sampling period, was first promoted in estimation and control applications (see, [12], [13]). Later on, the numerical properties of the delta operator, where $T_s$ is replaced by a positive factor $\Delta$, were investigated in [4] from a pure algebraic point of view. It was found that one can make the transfer function in delta operator have better numerical properties in the case where the poles of the transfer function are closer to $z = +1$ than $z = 0$. This means that the delta operator based structures have a very good performance for narrow band low-pass filters and may not yield a satisfactory performance for other types of filters. In this paper, our contribution is twofold. First of all, based on the concept of polynomial operators a more generalized DFIIt structure is derived. This structure is optimized with respect to the polynomial operators to reduce the roundoff noise gain. The second one is to study its equivalent state-space realization. It is shown that this realization always yields a smaller roundoff noise than that of the generalized DFIIt structure.

[1] By trivial parameters we mean those that are 0 and $\pm 1$. Other parameters are, therefore, referred to non-trivial parameters.

## 2. A POLYNOMIAL OPERATOR BASED DFIIT STRUCTURE

Consider the following time-invariant linear digital filter $H(z)$ given by

$$H(z) = \frac{b_0 z^p + b_1 z^{p-1} + ... + b_{p-1} z + b_p}{z^p + a_1 z^{p-1} + ... + a_{p-1} + a_p}. \quad (1)$$

Define

$$\rho_k \triangleq \frac{z - \gamma_k}{\Delta_k}, \; k = 1, 2, ..., p, \quad (2)$$

where $\{\gamma_k\}$ and $\{\Delta_k > 0\}$ are two sets of constants to be discussed later.

It can be shown that $H(z)$ can be reparametrized with $\{\alpha_m, \beta_m\}$ in the polynomials $\{\rho_k\}$, called polynomial operators:

$$H(z) = \frac{\beta_0 + \beta_1 \rho_1^{-1} + ... + +\beta_p \prod_{k=1}^{p} \rho_k^{-1}}{1 + \alpha_1 \rho_1^{-1} + ... + \alpha_p \prod_{k=1}^{p} \rho_k^{-1}}. \quad (3)$$

Denoting

$$\begin{aligned} V_a &\triangleq \begin{pmatrix} 1 & \cdots & a_p \end{pmatrix}^{\mathcal{T}}, \; V_b \triangleq \begin{pmatrix} b_0 & \cdots & b_p \end{pmatrix}^{\mathcal{T}} \\ V_\alpha &\triangleq \begin{pmatrix} 1 & \cdots & \alpha_p \end{pmatrix}^{\mathcal{T}}, \; V_\beta \triangleq \begin{pmatrix} \beta_0 & \cdots & \beta_p \end{pmatrix}^{\mathcal{T}}, \end{aligned}$$

one has

$$\begin{cases} V_a = \mathcal{K} M V_\alpha \\ V_b = \mathcal{K} M V_\beta \end{cases} \quad or \quad \begin{cases} V_\alpha = \mathcal{K}^{-1} M^{-1} V_a \\ V_\beta = \mathcal{K}^{-1} M^{-1} V_b \end{cases}, \quad (4)$$

where $\mathcal{K} = \prod_{k=1}^{p} \Delta_k$ and $M \in R^{(p+1)\times(p+1)}$ is a lower triangular matrix whose $mth$ column is determined by the coefficients of the polynomial $\prod_{k=m}^{p} \rho_k$ for $m = 1, 2, ..., p$ and $M(p+1, p+1) = 1$.

It can be shown that the output can be computed with the following equations

$$\begin{aligned} y(n) &= \beta_0 u(n) + w_1(n) \\ w_k(n) &= \rho_k^{-1}[\beta_k u(n) - \alpha_k y(n) + w_{k+1}(n)] \\ w_p(n) &= \rho_p^{-1}[\beta_p u(n) - \alpha_p y(n)] \end{aligned} \quad (5)$$

with $w_{p+1}(n) = 0$. Fig. 1 shows the corresponding realization structure to (5), where $w_k(n)$ is the output of the operator $\rho_k^{-1}$.

Clearly, when $\gamma_k = 0$, $\Delta_k = 1$, $\forall k$, Fig. 1 is the conventional DFIIt, and when $\gamma_k = 1$, $\forall k$, one gets the generalized $\delta$DFIIt structure, denoted as WN's structure, studied in [7]. In this paper, we just consider the cases for which $\gamma_k$ takes values from the set $\{-1, 0, 1\}$.

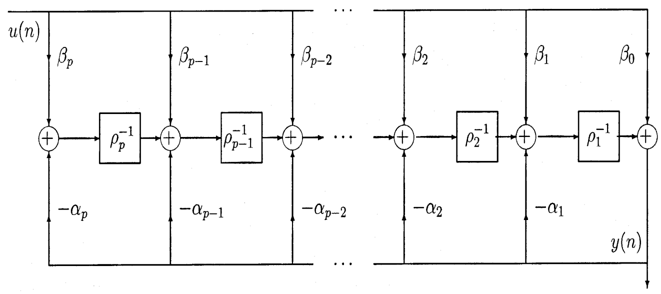


Fig. 1. A generalized DFIIt structure with polynomial operators

### 2.1. Equivalent state-space realization

One can implement $\rho_k^{-1}$ with the realization depicted in Fig. 2. We choose $\{x_k(n)\}$ indicated in Fig. 2 as the state variables and denote $x(n)$ as the state vector. It can be shown that the proposed structure is equivalent to the following state-space realization

$$\begin{aligned} x(n+1) &= A_\rho x(n) + B_\rho u(n) \\ y(n) &= C_\rho x(n) + \beta_0 u(n), \end{aligned} \quad (6)$$

where

$$\begin{aligned} B_\rho &= \bar{\beta} - \beta_0 \bar{\alpha}, \; C_\rho = \begin{pmatrix} \Delta_1 & 0 & \cdots & 0 \end{pmatrix} \\ A_\rho &= D_\gamma + M_\alpha \end{aligned} \quad (7)$$

with $D_\gamma = diag(\gamma_1, \cdots, \gamma_p)$, $M_\alpha$ is the $p \times p$ zero matrix except $M(k, 1) = -\Delta_1 \alpha_k, \forall k$ and $M_\alpha(k, k+1) = \Delta_{k+1}$ for $k = 1, \cdots, p-1$, and

$$\bar{\beta} \triangleq \begin{pmatrix} \beta_1 & \cdots & \beta_p \end{pmatrix}^{\mathcal{T}}, \; \bar{\alpha} \triangleq \begin{pmatrix} \alpha_1 & \cdots & \alpha_p \end{pmatrix}^{\mathcal{T}}.$$

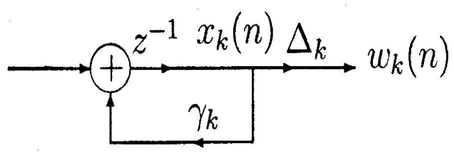


Fig. 2. A realization of operator $\rho_k^{-1}$

Denote $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$ as the realization given in (7) but corresponding to $\Delta_k = 1$, $\forall k$. It can be shown that the realization $(A_\rho, B_\rho, C_\rho, \beta_0)$ with any given $\{\Delta_k\}$ can be obtained from $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$ with a diagonal similarity transformation, denoted as $T_{sc}$,

$$A_\rho = T_{sc} \bar{A}_\rho T_{sc}^{-1}, \; B_\rho = T_{sc} \bar{B}_\rho, \; C_\rho = \bar{C}_\rho T_{sc}^{-1}, \quad (8)$$

where

$$T_{sc} = diag(d_1, d_2, \cdots, d_p), \quad d_k = \prod_{m=1}^{k} \Delta_m^{-1}, \; \forall k. \quad (9)$$

### 2.2. $l_2$-scaling

The $l_2$-scaling means that each state variable should have a unit variance when the input is a white noise with a unit variance. This can be achieved if

$$W_c(k,k) = 1, \forall k, \tag{10}$$

where $W_c$ is given by

$$W_c = \sum_{k=0}^{+\infty} A_\rho^k B_\rho B_\rho^{\mathcal{T}} (A_\rho^{\mathcal{T}})^k, \tag{11}$$

called the controllability gramian of $(A_\rho, B_\rho, C_\rho, \beta_0)$.

Denote $\bar{W}_c$, corresponding to $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$. It follows from (8)-(9) that $W_c = T_{sc}\bar{W}_c T_{sc}^{\mathcal{T}}$. Therefore, the $l_2$-scaling can be achieved if $d_k^2 \bar{W}_c(k,k) = 1$, $\forall k$, which leads to

$$\Delta_1 = \sqrt{\bar{W}_c(1,1)}, \; \Delta_k = \sqrt{\frac{\bar{W}_c(k,k)}{\bar{W}_c(k-1,k-1)}} \quad k = 2,3,...,p. \tag{12}$$

We note that for a given digital filter, one can implement it with the generalized DFIIt structure depicted in Fig. 1 as well as its equivalent state-space (SS) realization (6). In the next section, we will analyze the performance of the two structures in terms of roundoff noise gain.

## 3. ROUNDOFF NOISE ANALYSIS

Roundoff noise occurs in those variables computed with multiplications if less-than-double precision fixed-point arithmetic and rounding are utilized. Assuming rounding occurs after multiplication, expressions for roundoff noise gain in the above two structures are derived below.

### 3.1. Roundoff noise of the generalized DFIIt

Since $\{\gamma_k\}$ are trivial parameters, there is no rounding occuring after them at all. Therefore, the expression for the roundoff noise gain is exactly the same as that derived for WN's structure in [7] for a given set $\{\gamma_k\}$.

Denoting $W_o$ as the observability of the realization $(A_\rho,\, B_\rho, C_\rho, \beta_0)$, which is given by

$$W_o = \sum_{k=0}^{+\infty} (A_\rho^{\mathcal{T}})^k C_\rho^{\mathcal{T}} C_\rho A_\rho^k. \tag{13}$$

Denote $G_\alpha$, $G_\beta$, and $G_\Delta$ as the roundoff noise gain due to $\{\alpha_k\}, \{\beta_k\}$ and $\{\Delta_k\}$, respectively, the total roundoff noise gain is given by

$$G = 3tr(W_o) + 2(1 + \bar{\alpha}^{\mathcal{T}} W_o \bar{\alpha}) - W_o(p,p). \tag{14}$$

When $\gamma_k = 1, \forall k$, $G$ above yields to roundoff noise gain for the WN's structure [7].

### 3.2. Roundoff noise of the equivalent SS realization

First of all, we note that $A_\rho$ has two non-trivial parameters $A_\rho(k,1)$ and $A_\rho(k,k+1)$ in $kth$ row for $k = 1,2,...,p-1$ and one non-trivial parameter $A_\rho(p,1)$ in the $pth$ row, all the elements of $B_\rho$ are generally non-trivial, and $C_\rho$ has only one non-trivial parameter. And assume that the direct term $\beta_0$ is non-trivial. Noting that each of those parameters will produce a roundoff noise source after multiplication, (6) becomes

$$\begin{aligned} x^*(n+1) &= A_\rho x^*(n) + B_\rho u(n) + \eta(n) \\ y^*(n) &= C_\rho x^*(n) + \beta_0 u(n) + \varphi(n) \end{aligned} \tag{15}$$

where $\eta(n)$ is a noise vector whose $kth$ element is the summation of three and two roundoff noise sources for $k = 1,2,...,p-1$ and $k = p$, respectively, while $\varphi(n) = \epsilon_1^c(n) + \epsilon^{\beta_0}(n)$, where the two roundoff noise sources are due to $C_\rho(1)$ and $\beta_0$, respectively.

Denoting

$$e_x(n) \triangleq x(n) - x^*(n), \; e_y(n) \triangleq y(n) - y^*(n),$$

it can be shown that

$$\begin{aligned} e_y(n) &= -C_\rho(zI-A)^{-1}\eta(n) - \varphi(n) \\ &\triangleq \eta_e(n) - \varphi(n) \end{aligned} \tag{16}$$

Since all the roundoff noise sources are statistically independent zero-mean processes with the same variance $\sigma_0^2$, we have

$$\sigma^2 = E[e_y^2(n)] = E[\eta_e^2(n)] + E[\varphi^2(n)]. \tag{17}$$

With some manipulations, the roundoff noise gain, denoted as $G_{ss} \triangleq \frac{\sigma^2}{\sigma_0}$, can be shown to be given by the following

$$G_{ss} = tr(QW_o) + 2 = 3tr(W_o) - W_o(p,p) + 2, \tag{18}$$

where $Q = diag(3,3,\cdots,3,2)$ and $W_o$is given by (13).

Remark 3.1: Comparing (14) with (18), one can conclude that for the same $\{\gamma_k\}$, $G_{ss} < G$ is always true.

### 3.3. Optimal operators

As mentioned before, $\gamma_k$ takes value from the finite set $S_\gamma$:

$$S_\gamma \triangleq \{-1,\; 0,\; 1\}. \tag{19}$$

For a given $\{\gamma_k\}$, one can compute $(\bar{A}_\rho, \bar{B}_\rho, \bar{C}_\rho, \beta_0)$ and hence $(\bar{W}_c, \bar{W}_o)$. Then, the scaling factors $\{\Delta_k\}$ can be computed with either (12) and hence $T_{sc}$ can be determined with (9). It can be shown that $W_o = T_{sc}^{-1}\bar{W}_o T_{sc}^{-1}$, then the total roundoff noise gains $G$ and $G_{ss}$ can be evaluated with (14) and (18), respectively, for the generalized DFIIt structure and its equivalent state-space realization. We have the following two interesting problems

$$\min_{\gamma_k \in S_\gamma,\ \forall k} G, \qquad \min_{\gamma_k \in S_\gamma,\ \forall k} G_{ss}. \tag{20}$$

The first one leads to the optimal generalized DFIIt structure, while the second one, to the optimal equivalent state-space realization. Though $G$ and $G_{st}$ are highly nonlinear function of $\{\gamma_k\}$, the problem can be solved easily since the space $\{\gamma_k :\ \gamma_k \in S_\gamma\}$ is finite.

## 4. DESIGN EXAMPLES

This is a fourth order low-pass Butterworth filter with a normalized $3dB$ frequency $f_c = 0.125$. The corresponding poles are located at $p_{1,2} = 0.5565 \pm j0.5142$, $p_{3,4} = 0.4277 \pm j0.1637$ with $|p_{1,2}| = 0.7577$, $|p_{3,4}| = 0.4576$.

The following table shows the roundoff noise gains $G$ and $G_{ss}$ for five different sets of $\{\gamma_k\}$.

Table Roundoff Noise Gain

| $\gamma_1$ | $\gamma_2$ | $\gamma_3$ | $\gamma_4$ | $G$ | $G_{ss}$ |
|---|---|---|---|---|---|
| -1 | -1 | -1 | -1 | $5.3056 \times 10^3$ | $4.0095 \times 10^3$ |
| 0 | 0 | 0 | 0 | 52.4953 | 29.0450 |
| 1 | 1 | 1 | 1 | 27.1189 | 22.5630 |
| 0 | 1 | 0 | 1 | 13.9221 | 10.3263 |
| 0 | 0 | 0 | 1 | 14.8786 | 8.7249 |

Here, we note that $\{0,\ 0,\ 0,\ 0\}$ and $\{1,\ 1,\ 1,\ 1\}$ correspond to the zDFIIt (in the shift operator) and the $\delta$DFIIt (in the delta operator), as proposed in [7], respectively. The combinations $\bar{\gamma}_G \triangleq \{0,\ 1,\ 0,\ 1\}$ and $\bar{\gamma}_{G_{ss}} \triangleq \{0,\ 0,\ 0,\ 1\}$ yield the optimal polynomial operators obtained from (20) for $G$ and $G_{ss}$, respectively. We can see that for the generalized DFIIt the optimal polynomial operators $\bar{\gamma}_G$ yields a $G$ which is just half of that by $\delta$DFIIt. And for optimal SS realization $\bar{\gamma}_{G_{ss}}$, it is better than $\bar{\gamma}_G$ and has a much smaller roundoff noise gain, just one third of that yielded by $\delta$DFIIt.

## 5. REFERENCES


[1] S.Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," IEEE Trans. on Acoust., Speech, and Signal Processing, Vol. ASSP-25, No. 4, pp. 273-281, Aug. 1977.

[2] L. Thiele, "On the sensitivity of linear state-space systems," IEEE Trans. on Circuits and Systems, Vol. CAS-33, pp. 502-510, May 1986.

[3] R.A. Roberts and C.T. Mullis, Digital Signal Processing, Reading, MA: Addison Wesley, 1987.

[4] Michel Gevers and Gang Li, Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects, Springer Verlag London, Communication and Control Engineering Series, 1993.

[5] G. Amit and U. Shaked, "Small roundoff realization of fixed-point digital filters and controllers," IEEE on Trans. on Acoust., Speech, and Signal Processing, Vol. ASSP-36, no. 6, pp. 880-891, Jun. 1988.

[6] G. Li, M. Gevers, and Y.X. Sun, "Performance analysis of a new structure for digital filter implementation," IEEE on Trans. Circuits and Systems-I, Vol. 47, pp. 474-482, April 2000.

[7] N. Wong and T.S. Ng, "A generalized direct-form delta operator-based IIR filter with minimum noise gain and sensitivity," IEEE Trans. on Circuits and Systems - II, Vol. 48, pp. 425-431, April 2001.

[8] N. Wong and T.S. Ng, "Roundoff noise minimization in a modified direct form delta operator IIR structure," IEEE Trans. on Circuits and Systems - II, Vol. 47, pp. 1533-1536, Dec. 2000.

[9] J. Kauraniemi and T.I. Laakso, "Roundoff noise analysis of modified delta operator direct form structures," in Proc. IEEE Int. Symp. Circuits Syst., 1997.

[10] J. Kauraniemi, T.I. Laakso, I. Harttimo, and S.J. Ovaska, "Roundoff noise minimization in a direct form delta operator structure," in Proc. IEEE Int. Acoust., Speech, and Signal Processing, Atlanta, GA, May, 1996.

[11] J. Kauraniemi, T.I. Laakso, I. Harttimo, and S.J. Ovaska , "Delta operator realizations of direct-form IIR filters," IEEE Trans. Circuits and Systems-II, Vol. 45, pp. 41-45, Jan. 1998.

[12] V. Peterka, "Control of uncertain processes: applied theory and algorithms," Kybernetika, Vol. 22, pp. 1-102, 1986.

[13] R.H. Middleton and G.C. Goodwin, Digital Control and Estimation: A Unified Approach, Prentice Hall, Englewood Cliffs, New Jersey, 1990.